

Scaling Behaviors of Wireless Device-to-Device Communications with Distributed Caching

Negin Golrezaei, Alexandros G. Dimakis, Andreas F. Molisch

Dept. of Electrical Eng.

University of Southern California

emails: {golrezae,dimakis,molisch}@usc.edu

Abstract—We analyze a novel architecture for caching popular video content to enable wireless device-to-device collaboration. We focus on the asymptotic scaling characteristics and show how they depend on video content popularity statistics. We identify a fundamental conflict between collaboration distance and interference and show how to optimize the transmission power to maximize frequency reuse.

Our main result is a closed form expression of the optimal collaboration distance as a function of the model parameters. Under the common assumption of a Zipf distribution for content reuse, we show that if the Zipf exponent is greater than 1, it is possible to have a number of D2D interference-free collaboration pairs that scales linearly in the number of nodes. If the Zipf exponent is smaller than 1, we identify the best possible scaling in the number of D2D collaborating links. Surprisingly, a very simple distributed caching policy achieves the optimal scaling behavior and therefore there is no need to centrally coordinate what each node is caching.

I. INTRODUCTION

Wireless mobile data traffic is expected to increase by a factor of 40 over the next five years, from the current 93 Petabytes to 3600 Petabytes per month in the next five years [1]. This explosive demand is fueled mainly by mobile video traffic that is expected to increase by a factor of 65, and become the by far dominant source of data traffic. Since the available spectrum is physically limited and the spectral efficiency of current systems is already close to optimum, the main method for meeting this increased demand is to bring content closer to the users. Femto base stations [2] are currently receiving a lot of attention for this purpose.

A significant bottleneck in such small-cell architectures is that each station requires a high-rate backhaul link. Helper stations that replace high-rate backhaul with storage [3] [4], can ameliorate the problem, but still require additional infrastructure and have limited flexibility.

To circumvent these problems, we recently proposed the use of device-to-device (D2D) communications combined with video caching in mobile devices [5] [7]. The approach is based on three key observations: (i) Modern smartphones and tablets have significant storage capacity, (ii) video has a large amount of *content reuse*, i.e., a small number of video files accounts for a large fraction of the traffic. (iii) D2D communication can occur over very short distances thus allowing high frequency reuse. Our proposed architecture functions as follows: users can collaborate by caching popular content and utilizing local D2D communication when a user in the vicinity requests a popular file. The base station can keep track of the availability of the cached content and direct requests to the most suitable nearby device; if there is no suitable nearby device, the BS

supplies the requested video file directly, via a traditional downlink transmission. Storage allows users to collaborate even when they do not request the same content *at the same time*. This is a new dimension in wireless collaboration architectures beyond relaying and cooperative communications as in [6] [5] and references therein.

A D2D video network can be analyzed using a protocol model, which means that only two devices that are within a "collaboration distance" of each other can exchange video files, while devices with a larger distance do not create any useful signal, but also no interference, for each other. The choice of the collaboration distance represents a tradeoff between two counteracting effects: decreasing the collaboration distance increases the frequency reuse and thus the potential throughput, but on the other hand decreases the probability that a device can find a requested file cached on another device within the collaboration distance. In [7] we described this tradeoff and provided numerical solutions for the optimum distance, and the resulting system throughput.

In the current paper we concentrate on the analytical treatment of the *scaling behavior* of a D2D network, i.e., how the throughput scales as the number of nodes increases. For conventional ad-hoc networks, scaling behavior has been derived in the seminal paper by Gupta and Kumar [8] has further received significant attention (e.g. see [9]–[11]). This architecture not only differs from ad-hoc or collaborative networks in its application, but also shows a fundamentally different behavior due to its dependence on the video reuse statistics. We provide a closed form expression of the optimal collaboration distance as a function of the content reuse distribution parameters.

We model the request statistics for video files by a Zipf distribution which has been shown to fit well with measured YouTube video requests [12] [13]. We find that the scaling laws depend critically on the Zipf parameter, i.e., on the concentration of the request distribution. We show that if the Zipf exponent of the content reuse distribution is greater than 1, it is possible to have a number of D2D interference-free collaboration pairs that scales linearly with the number of nodes. If the Zipf exponent is smaller than 1, we identify the best possible scaling in the number of D2D collaborating links. Surprisingly, a very simple distributed caching policy achieves the optimal scaling behavior and therefore there is no need to centrally coordinate what each node is caching. For Zipf exponent equal to 1, we find the best collaboration distance and the best possible scaling.

The remainder of this paper is organized as follows: In Sec-

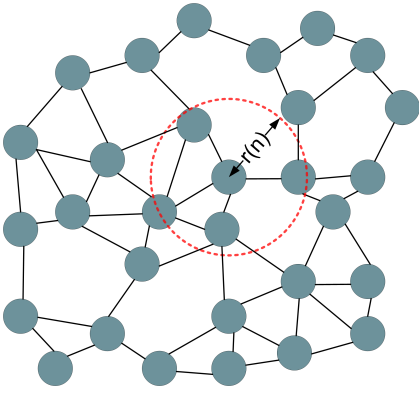


Fig. 1. Random geometric graph example with collaboration distance $r(n)$.

tion II we set up the D2D formulation and explain the tradeoff between collaboration distance and interference. Section III contains our two main theorems, the scaling behavior for Zipf exponents greater, smaller than and equal to 1. In Section IV we discuss future directions, open problems and conclusions. Finally, the Appendix contain the proofs of our theorems.

II. MODEL AND SETUP

In this section, we discuss the fundamental system model; for a discussion of the assumptions, and justifications of simplifications, we refer the interested reader to [7].

Assume a cellular network where each cell/base station (BS) serves n users. For simplicity we assume that the cells are square, and we neglect inter-cell interference, so that we can consider one cell in isolation. Users are distributed randomly and independently in the cell. We assume that the D2D communication does not interfere with the base station that can serve video requests that cannot be otherwise covered. For that reason, our only concern is the maximization of the number of D2D collaboration links that can be simultaneously scheduled. We henceforth do not need to consider explicitly the BS and its associated communications.

The communication is modeled by a standard protocol model on a random geometric graph (RGG) $G(n, r(n))$. In this model users are randomly and uniformly distributed in a square (cell) of size 1. Two users (assuming D2D communication is possible) can communicate if their euclidean distance is smaller than some collaboration distance $r(n)$ [8], [14]. The maximum allowable distance for D2D communication $r(n)$ is determined by the power level for each transmission. Figure 1 illustrates an example of an RGG.

We assume that users may request files from a set of size m that we call a “library”. The size of this set should increase as a function of the number of users n . Intuitively, the set of YouTube videos requested in Berkeley in one day should be smaller than the set of requested in Los Angeles. We assume that this growth should be sublinear in n , e.g. m could be

$\Theta(\log(n))$ ¹.

Each user requests a file from the library by sampling independently using a popularity distribution. Based on several studies, Zipf distributions have been established as good models for the measured popularity of video files [12], [13]. Under this model, the popularity of the i th popular file, denoted by f_i , is inversely proportional to its rank:

$$f_i = \frac{1}{i^{\gamma_r} \sum_{j=1}^m \frac{1}{j^{\gamma_r}}}, \quad 1 \leq i \leq m. \quad (1)$$

The Zipf exponent γ_r characterizes the distribution by controlling the relative popularity of files. Larger γ_r exponents correspond to higher content reuse, i.e., the first few popular files account for the majority of requests.

Each user has a storage capacity called cache which is populated with some video files. For our scaling law analysis we assume that all files have the same size, and each user can store one file. This yields a clean formulation and can be easily extended for larger storage capacities.

Our scheme works as follows: If a user requests one of the files stored in neighbors’ caches in the RGG, neighbors will handle the request locally through D2D communication; otherwise, the BS should serve the request. Thus, to have D2D communication it is not sufficient that the distance between two users be less than $r(n)$; users should find their desired files locally in caches of their neighbors. A link between two users will be called potentially *active* if one requests a file that the other is caching. Therefore, the probability of D2D collaboration opportunities depends on what is stored and requested by the users.

The decision of what to store can be taken in a distributed or centralized way. A central control of the caching by the BS allows very efficient file-assignment to the users. However, if such control is not desired or the users are highly mobile, caching has to be optimized in a distributed way. The simple randomized caching policy we investigate makes each user choose which file to cache by sampling from a caching distribution. It is clear that popular files should be stored with a higher probability, but the question is how much redundancy we want to have in our distributed cache.

We assume that all D2D links share the same time-frequency transmission resource within one cell area. This is possible since the distance between requesting user and user with the stored file will typically small. However, there should be no interference of a transmission by others on an active D2D link. We assume that (given that node u wants to transmit to node v) any transmission within range $r(n)$ from v (the receiver) can introduce interference for the $u - v$ transmission. Thus, they cannot be activated simultaneously. This model is known as *protocol model*; while it neglects important wireless propagation effects such as fading [15], it can provide fundamental

¹We use the standard Landau notation: $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$ respectively denote $|f(n)| \leq c_1 g(n)$ and $|f(n)| \geq c_2 g(n)$ for some constants c_1, c_2 . $f(n) = \Theta(g(n))$, stands for $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$. Little-o notation, i.e., $f(n) = o(g(n))$ is equivalent to $\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0$.

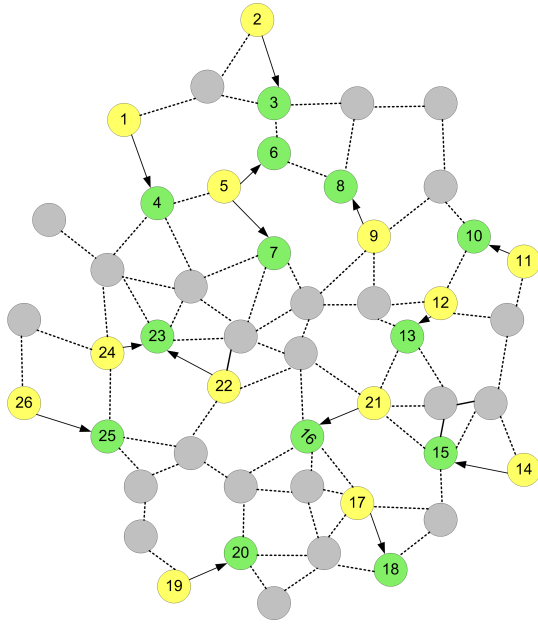


Fig. 2. Random geometric graph, yellow and green nodes indicate receivers, transmitters in D2D links. Gray nodes get their request files from the BS. Arrows show all possible D2D links.

insights and has been widely used in prior literature [8].

To model interference given a storage configuration and user requests we start with all potential D2D collaboration links. Then, we construct the conflict graph as follows. We model any possible D2D link between node u as transmitter to node v as a receiver with a vertex $u-v$ in the conflict graph. Then, we draw an edge between any two vertices (links) that create interference for each other according to the protocol model. Figure 3 shows how the RGG in Figure 2 is converted to the conflict graph. In Figure 2, receiver nodes are green and transmitter nodes are yellow. The nodes that should receive their desired files from the BS are gray. A set of D2D links is called active if they are potentially active and can be scheduled simultaneously, *i.e.*, form an independent set in the conflict graph. The random variable counting the number of active D2D links under some policy is denoted by L .

Figure 3 shows the conflict graph and one of maximum independent sets for the conflict graph. We can see that out of 14 possible D2D links 9 links can co-exist without interference. As is well known, determining the maximum independent set of an arbitrary graph is computationally intractable (NP complete [16]). Despite the difficulty of characterizing the number of interference-free active links, we can determine the best possible scaling law in our random ensemble.

III. ANALYSIS

A. Finding the optimal collaboration distance

We are interested in determining the best collaboration distance $r(n)$ and caching policy such that the expected number of active D2D links is maximized. Our optimization is based on balancing the following tension: The smaller the transmit power, the smaller the region in which a D2D communication creates interference. Therefore, more D2D pairs

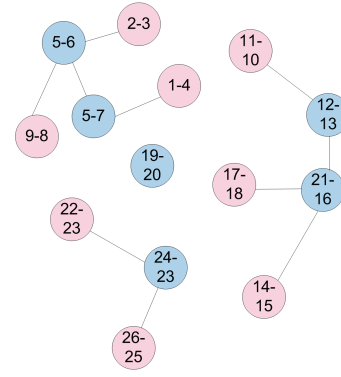


Fig. 3. conflict graph based on Figure 2 and one of maximum independent set of the conflict graph; pink vertices are those D2D links that can be activated simultaneously.

can be packed into the same area allowing higher frequency reuse. On the other hand, a small transmit power might not be sufficient to reach a mobile that stores the desired file. Smaller power means smaller distance and hence smaller probability of collaboration opportunities.

We analyze the case where the nodes do not possess power control with fast adaptation, but rather all users have the same transmit power that depends only on the node density. We then show how to optimize it based on the content request statistics. Our analysis involves finding the best compromise between the number of possible parallel D2D links and the probability of finding the requested content, as discussed above. Our results consist of two parts. In the first part (upper bound), we find the best achievable scaling for the expected number of active D2D links. In the second part (achievability), we determine an optimal caching policy and $r(n)$ to obtain the best scaling for the expected number of active links $E[L]$.

The best achievable scaling for the expected number of active D2D links depends on the extend of content reuse. Larger Zipf distribution exponents correspond to more redundancy in the user requests and a small number of files accounts for the majority of video traffic. Thus, the probability of finding requested files through D2D links increases by having access to few popular files via neighbors.

We separate the problem into three different regions depending on the Zipf exponent: $\gamma_r > 1$, $\gamma_r < 1$, and $\gamma_r = 1$. For each of these regions, we find the best achievable scaling for $E[L]$ and the optimum asymptotic $r(n)$ denoted by $r_{opt}(n)$. We also show that for $\gamma_r > 1$ and $\gamma_r < 1$ regions a simple distributed caching policy has optimal scaling, *i.e.*, matches the scaling behavior that any centralized caching policy could achieve. This caching policy means that each device stores files randomly, with a properly chosen caching distribution, namely a Zipf distribution with parameter γ_c . For $\gamma_r = 1$, we present an optimal centralized caching policy.

Our first result is the following theorem:

Theorem 1: If the Zipf exponent $\gamma_r > 1$,

- i) **Upper bound:** For any caching policy, $E[L] = O(n)$,
- ii) **Achievability:** Given that $c_1 \sqrt{\frac{1}{n}} \leq r_{opt}(n) \leq c_2 \sqrt{\frac{1}{n}}$

² and using a Zipf caching distribution with exponent $\gamma_c > 1$ then $E[L] = \Theta(n)$.

The first part of the theorem 1 is trivial since the number of active D2D links can at most scale linearly in the number of users. The second part indicates that if we choose $r_{opt}(n) = \Theta(\sqrt{\frac{1}{n}})$ and $\gamma_c > 1$, $E[L]$ can grow linearly with n . There is some simple intuition behind this result: We show that in this regime users are surrounded by a constant number of users in expectation. If the Zipf exponent γ_c is greater than one, this suffices to show that the probability that they can find their desired files locally is a non-vanishing constant as n grows. Our proof is provided in the Appendix A.

For the low content reuse region $\gamma_r < 1$, we obtain the following result:

Theorem 2: If $\gamma_r < 1$,

- i) **Upper bound:** For any caching policy, $E[L] = O(\frac{n}{m^\eta})$ where $\eta = \frac{1-\gamma_r}{2-\gamma_r}$,
- ii) **Achievability:** If $c_3 \sqrt{\frac{m^{\eta+\epsilon}}{n}} \leq r_{opt}(n) \leq c_4 \sqrt{\frac{m^{\eta+\epsilon}}{n}}$ and users cache files randomly and independently according to a Zipf distribution with exponent γ_c , for any exponent $\eta + \epsilon$, there exists γ_c such that $E[L] = \Theta(\frac{n}{m^{\eta+\epsilon}})$ where $0 < \epsilon < \frac{1}{6}$ and γ_c is a solution to the following equation

$$\frac{(1-\gamma_r)\gamma_c}{1-\gamma_r+\gamma_c} = \eta + \epsilon.$$

Our proof is provided in the Appendix B.

We show that when there is low content reuse, linear scaling in frequency re-use is not possible. At a high level, in order to achieve the optimal scaling, on average a user should be surrounded by $\Theta(m^\eta)$ users. Comparing with the first region where $\gamma_r > 1$, we can conclude that when there is less redundancy, users have to see more users in the neighborhood to find their desired files locally.

Theorem 3: If $\gamma_r = 1$

- i) **Upper bound:** For any $r(n)$, $E[L] = O(\frac{n \log \log(m)}{\log(m)})$
- ii) **Achievability:** Given that $c_5 \sqrt{\frac{\log(m)}{n \log \log(m)}} \leq r(n) \leq c_6 \sqrt{\frac{\log(m)}{n \log \log(m)}}$, there exists a centralized strategy such that

$$E[L] = \Theta(\frac{n \log \log(m)}{\log(m)}).$$

IV. DISCUSSION AND CONCLUSIONS

As mentioned in Sec. I, the study of scaling laws of the capacity of wireless networks has received significant attention since the pioneering work by Gupta and Kumar [8] (e.g. see [9]–[11]). The first result was pessimistic: if n nodes are trying to communicate (say by forming $n/2$ pairs), since the typical distance in a 2D random network will involve roughly $\Theta(\sqrt{n})$ hops, the throughput per node must vanish, approximately scaling as $1/\sqrt{n}$. There are, of course, sophisticated arguments performing rigorous analysis that sharpens the bounds and numerous interesting model extensions. One that is particularly relevant to this project is the work by Grossglauser and

Tse [10] that showed that if the nodes have infinite storage capacity, full mobility and there is no concern about delay, constant (non-vanishing) throughput per node can be sustained as the network scales.

Despite the significant amount of work on ad hoc networks, there has been very little work on file sharing and content distribution over wireless ([3], [17]) beyond the multiple unicast traffic patterns introduced in [8]. Our result shows that if there is sufficient content reuse, caching fundamentally changes the picture: non-vanishing throughput per node can be achieved, even with constant storage and delay, and without any mobility.

On a more technical note, the most surprising result is perhaps the fact that in Theorem 2, a simple distributed policy can match the optimal scaling behavior $E[L] = O(\frac{n}{m^\eta})$. This means that even if it were possible for a central controller to impose on the devices what to store, the scaling behavior could not improve beyond the random caching policy (though, of course, the actual numerical values for finite device density could be different). Further, for both regimes of γ_r , the distributed caching policy exponent γ_c should not match the request Zipf exponent γ_r , something that we found quite counter intuitive.

Overall, even if linear frequency re-use is not possible, we expect the scaling of the library m to be quite small (typically logarithmic) in the number of users n . In this case we obtain near-linear (up to logarithmic factors) growth in the number of D2D links for the full spectrum of Zipf exponents. Our results are encouraging and show that device-based caching and D2D communications can lead to drastic increase of wireless video throughput; and that the benefits increase as the number of participants increases. This in turn implies that the highest throughput gains are achieved in those areas where they are most needed, i.e., where the devices are most concentrated.

APPENDIX A PROOF OF THEOREM 1

The first part of the theorem is easy to see since the number of D2D links cannot exceed the number of users. Next, we show the second part of the theorem.

For the second part of the theorem, we introduce virtual clusters and we show that the number of virtual clusters that can be potentially active, called *good clusters*, scales like the number of active links. To find the lower bound for good clusters, we limit users to communicate with neighbors in the same cluster. Then, we express the probability of good cluster as function of stored files by users within the cluster. Excluding *self-requests*, i.e., when users find their request files in their own caches, we find a lower bound for good clusters. We further define a *value* for each cluster which is the sum of probability of stored files by users. Then we express the probability of goodness as a function of value of clusters. Using Chernoff bound, we finalize our proof.

A. Active links versus good clusters

We divide the cell into $\frac{2}{r(n)^2}$ virtual square clusters. Figure 4(a) shows the virtual clusters in the cell. The cell side is

² c and c_i s are positive constants that do not depend on n .

normalized to 1 and the side of each cluster is equal to $\frac{r(n)}{\sqrt{2}}$. Thus, all users within a cluster can communicate with each other. Based on our interference model, in each cluster only one link can be activated. When there is an active D2D link within a cluster, we call the cluster *good*. But not all good clusters can be activated simultaneously. According to protocol model, one good cluster can at most block 16 clusters (see Figure 4(b)). The maximum interference happens when a user in the corner of a cluster transmits a file to a user in the opposite corner. So, we have

$$E[L] \geq \frac{E[G]}{(16 + 1)} \quad (2)$$

where $E[G]$ is the expected number of good clusters.

Since the number of active links scales like the number of good clusters, to prove the theorem it is enough to show that constant fraction of virtual clusters are good. This is because $r(n) = \Theta(\sqrt{\frac{1}{n}})$ and there are $\Theta(n)$ virtual clusters in the cell.

B. Limiting users

Since we want to find the lower bound for $E[L]$, we can limit users to communicate with users in virtual clusters they belong to. Hence,

$$E[G] \geq \frac{2}{r(n)^2} \sum_{k=0}^n \Pr[\text{good}|k] \Pr[K = k], \quad (3)$$

where $\frac{2}{r(n)^2}$ is the total number of virtual clusters. K is the number of users in the cluster, which is a binomial random variable with n trials and probability of $\frac{r(n)^2}{2}$, i.e., $K = B(n, \frac{r(n)^2}{2})$. $\Pr[K = k]$ is the probability that there are k users in the cluster and $\Pr[\text{good}|k]$ is the probability that the cluster is good conditioned on k .

C. Probability of goodness and stored files

To show the result, we should prove that the summation in (3), i.e., the probability that a cluster is good, does not vanish as n goes to infinity. The probability that a cluster is good depends on what users cache. Therefore,

$$E[G] \geq \frac{2}{r(n)^2} \sum_{k=0}^n \Pr[K = k] \times \sum_{\{\omega \mid |\omega|=k\}} \Pr[\text{good}|k, \omega] \Pr[\omega], \quad (4)$$

where ω is a random vector of stored files by users in the cluster and $|\omega|$ denotes the length of vector ω . The i th element of ω denoted by $\omega_i \in \{1, 2, 3, \dots, m\}$ indicates what user i in the cluster stores.

For each ω , we define a value:

$$v(\omega) = \sum_{i \in \tilde{\omega}} f_i, \quad (5)$$

where $\tilde{\omega} = \bigcup_{j=1}^{|\omega|} \omega_j$ and \bigcup is the union operation. Actually $v(\omega)$ is the sum of popularities of the union of files in ω . The cluster is considered to be good if at least a user i in the cluster requests one of the files in $\tilde{\omega} - \{\omega_i\}$.

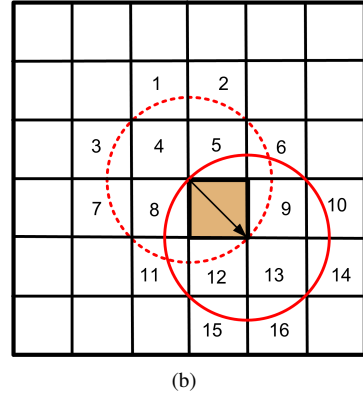
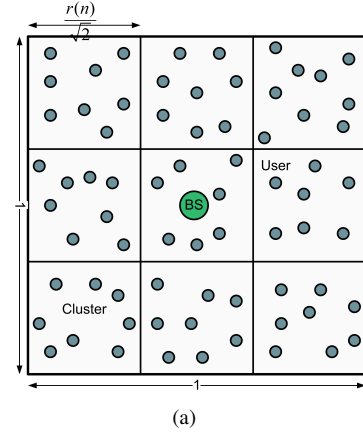


Fig. 4. a) Dividing cell into virtual clusters. b) In the worst case, a good cluster can block at most 16 clusters. In the dashed circle, receiving is not possible and in the solid circle, transmission is not allowed.

D. Excluding self request

A user might find the file it requests in its own cache; in this case clearly no D2D communication will be activated by this user. We call these cases *self-requests*. Accounting for these self-requests, the probability that user i finds its request files locally within the cluster is $(v(\omega) - f_{\omega_i})$. Thus, we obtain:

$$\Pr[\text{good}|k, \omega] \geq 1 - (1 - (v(\omega) - \max_i f_{\omega_i}))^k. \quad (6)$$

Let us only consider cases where at least one user in the cluster caches file 1 (the most popular file). Then, from (4) and (6), the following lower bound is achieved:

$$E[G] \geq \frac{2}{r(n)^2} \sum_{k=1}^n \Pr[K = k] \times \sum_{\omega \in \mathbf{x}} [1 - (1 - (v(\omega) - f_1))^k] \Pr[\omega], \quad (7)$$

where $\mathbf{x} = \{\omega \mid |\omega| = k \text{ and } 1 \in \tilde{\omega}\}$.

E. Probability of goodness and value of clusters

Instead of taking expectation with respect to ω , we take expectation with respect to v , i.e., the value of a cluster. Then,

$$E[G] \geq \frac{2}{r(n)^2} \sum_{k=1}^n \Pr[K = k] E_v[1 - (1 - (v - f_1))^k | A_1^k]$$

$$\geq \frac{2}{r(n)^2} \sum_{k=1}^n \Pr[K = k] E_v[(v - f_1) | A_1^k],$$

where A_1^k is the event that at least one of k users in the cluster caches file 1 and $E_v[\cdot]$ is the expectation with respect to v . Let $A_{1,h}^k$ for $1 \leq h \leq k$ denote the event that h users out of k users in the cluster cache file 1. Then, we get:

$$\begin{aligned} E[G] &\geq \frac{2}{r(n)^2} \sum_{k=1}^n \Pr[K = k] \\ &\quad \times \sum_{h=1}^k E_v[(v - f_1) | A_{1,h}^k] \times \Pr[A_{1,h}^k], \end{aligned} \quad (8)$$

where $\Pr[A_{1,h}^k] = \binom{k}{h} (p_1)^h (1 - p_1)^{k-h}$ and p_j represents the probability that file j is cached by a user based on Zipf distribution with exponent γ_c . To calculate $E_v[(v - f_1) | A_{1,h}^k]$, we define an indicator function $\mathbf{1}_j$ for each file $j \geq 2$. $\mathbf{1}_j$ is equal to 1 if at least one user in the cluster stores file j . Hence,

$$\begin{aligned} E_v[(v - f_1) | A_{1,h}^k] &= E\left[\sum_{j=2}^m f_j \mathbf{1}_j | A_{1,h}^k\right] \\ &= \sum_{j=2}^m f_j (1 - (1 - p_j)^{k-h}). \end{aligned}$$

F. Chernoff bound

To show that the probability of a cluster is good is not vanishing, we use Chernoff bound. First, we limit the interval k to an interval around its average. By substituting $E_v[(v - f_1) | A_{1,h}^k]$ in (8),

$$\begin{aligned} E[G] &\geq \frac{2}{r(n)^2} \sum_{k \in I} \Pr[K = k] \\ &\quad \times \sum_{h=1}^k \sum_{j=2}^m f_j (1 - (1 - p_j)^{k-h}) \Pr[A_{1,h}^k], \end{aligned} \quad (9)$$

where for any $0 < \delta < 1$ the interval $I = [nr(n)^2(1 - \delta)/2, nr(n)^2(1 + \delta)/2]$. Define $k^* \in I$ such that it minimizes the expression in the last line of (9). Since $r(n) = \Theta(\sqrt{\frac{1}{n}})$, k^* is $\Theta(1)$. Then from (9), we have:

$$\begin{aligned} E[G] &\geq \frac{2}{r(n)^2} \Pr[K \in I] \\ &\quad \times \sum_{h=1}^{k^*} \left[\Pr[A_{1,h}^{k^*}] \sum_{j=2}^m f_j (1 - (1 - p_j)^{k^*-h}) \right] \\ &\geq \frac{2}{r(n)^2} (1 - 2 \exp(-nr(n)^2 \delta^2/6)) \\ &\quad \times \sum_{h=k^* p_1(1+\delta_1)}^{k^* p_1(1+\delta_1)} \left[\Pr[A_{1,h}^{k^*}] \sum_{j=2}^m f_j (1 - (1 - p_j)^{k^*-h}) \right], \end{aligned} \quad (11)$$

where $0 < \delta_1 < 1$. We apply the Chernoff bound in (10) to derive (11) [18]. Since the exponent $nr(n)^2 \delta^2/6$

is $\Theta(1)$, we can select the constant c_1 such that the term $1 - 2 \exp(-nr(n)^2 \delta^2/6)$ becomes positive.

Let us define $h^* \in [k^* p_1(1 - \delta_1), k^* p_1(1 + \delta_1)]$ such that it minimizes the inner summation of (11), i.e., $\sum_{j=2}^m f_j (1 - (1 - p_j)^{k^*-h})$. From (1), p_1 is $\frac{1}{H(\gamma_c, 1, m)}$ where function H is defined in lemma 1 in Appendix. Some preliminary lemmas. Lemma 1 implies that $p_1 = \Theta(1)$ and as a result, h^* is also $\Theta(1)$. Using the Chernoff bound for random variable h in (11), we get:

$$\begin{aligned} E[G] &\geq \frac{2}{r(n)^2} (1 - 2 \exp(-nr(n)^2 \delta^2/6)) \\ &\quad \times (1 - 2 \exp(-k^* p_1 \delta_1^2/3)) \sum_{j=2}^m f_j (1 - (1 - p_j)^{k^*-h^*}). \end{aligned} \quad (12)$$

$k^* - h^*$ should be greater than 1 which results in a constant lower bound for c_1 . The second exponent, i.e., $k^* p_1 \delta_1^2/3$ is $\Theta(1)$. Therefore, the term $(1 - 2 \exp(-k^* p_1 \delta_1^2/3))$ is a positive constant if c_1 is large enough. Further, the summation in (12) satisfies

$$\sum_{j=2}^m f_j (1 - (1 - p_j)^{k^*-h^*}) > \sum_{j=2}^m f_j p_j.$$

To show that $E[G]$ scales linearly with n , the term $\sum_{j=2}^m f_j p_j$ should not be vanishing as n goes to infinity. Using part (iv) of lemma 1, we can see that if $\gamma_r, \gamma_c > 1$, $\sum_{j=2}^m f_j p_j = \Theta(1)$.

APPENDIX B PROOF OF THEOREM 2

To show the first part of the theorem, like the proof of theorem 1, we use virtual clusters. We show that the number active links can be at most equal to the number of good clusters. We state the probability of goodness as a function of stored files. To be more precise, we express this probability as an intersection of some decreasing events. Then, we use FKG inequality, to find an upper bound for probability of goodness. Finally, we divide the whole range of $r(n)$ into four non overlapping regions and show the upper bound for all regions.

A. Active links versus good clusters

To show the first part of the theorem, as in proof of the theorem 1, we divide the cell into $\frac{2}{r(n)^2}$ virtual square clusters. All users within a cluster can communicate with each other. Based on the protocol model, in each cluster only one link can be activated. As stated before, when there is an active D2D link within a cluster, we call the cluster good. In the best case, all the good clusters can be activated simultaneously. Hence,

$$E[L] \leq E[G],$$

where $E[G]$ is average number of good clusters. All users can look for their desired files not only in their own clusters but in the caches of all users in their vicinities. The maximum area that can be covered by all users in a cluster cannot be larger

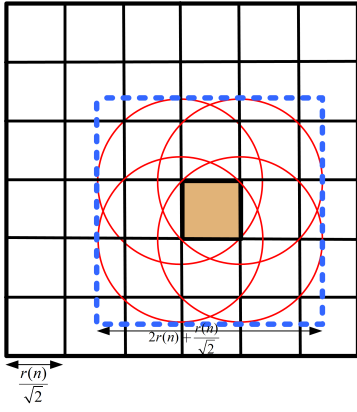


Fig. 5. Maximum area covered by all users within a cluster (blue square)

than $\alpha r(n)^2$ where $\alpha \triangleq (\frac{1}{\sqrt{2}} + 2)^2$ (the area of dashed square in Figure 5). Therefore,

$$E[L] \leq \frac{2}{r(n)^2} \sum_{k=0}^n \Pr[\text{good}|k] \Pr[K = k], \quad (13)$$

where K is the the number of users in dashed square (called *maximum square*) in Figure 5 which is binomial random variable with n trials and probability of $\alpha r(n)^2$, $K = B(n, \alpha r(n)^2)$.

B. Probability of goodness and stored files

$\Pr[\text{good}|k]$ is the probability that a cluster is good conditioned on k and it depends on what users in the maximum square stores denoted by ω .

$$\Pr[\text{good}|k] = \sum_{\{\omega \mid |\omega|=k\}} \Pr[\text{good}|k, \omega] \Pr[\omega]$$

Let's define an event $A_i(\omega)$ that user i finds its request either in the cache of its neighbors or its own cache.

$$\begin{aligned} \Pr[\text{good}|k, \omega] &\leq \Pr[A_1(\omega) \cup A_2(\omega) \cup \dots \cup A_k(\omega)] \\ &= 1 - \Pr[\bar{A}_1(\omega) \cap \bar{A}_2(\omega) \cap \dots \cap \bar{A}_k(\omega)] \end{aligned} \quad (14)$$

Events $A_i(\omega)$ and $A_j(\omega)$ for $j \neq i$ are dependent since they both depend on ω . The probability that event $A_i(\omega)$ happens is:

$$\Pr[A_i(\omega)] = \sum_{j=1}^m f_j \mathbf{1}_j,$$

where f_j is the probability that user i requests file j . $\mathbf{1}_j$ is an indicator function for file j and it is one if file $j \in \omega$. It is easy to check that $v(\omega)$ in (5) is equal to $\Pr[A_i(\omega)]$ for any i .

C. Increasing events and FKG inequality

To find an upper bound for intersection of dependent events $\bar{A}_i(\omega)$ s in (14), we first show that they are decreasing events. Then, we use the FKG inequality for decreasing events [19].

Definition 1: (Increasing event). A random variable X is increasing on (Ω, F) if $X(\omega) \leq X(\omega')$ whenever $\omega \leq \omega'$. It is decreasing if $-X$ is increasing.

We assume that $\omega \leq \omega'$ if the value of ω is less than the value of ω' , i.e.,

$$v(\omega) \leq v(\omega').$$

where the value of ω is defined in (5). Thus, according to this definition, event $A_i(\omega)$ for any $1 \leq i \leq k$ is an increasing event. Applying the FKG inequality for correlated and decreasing events $\bar{A}_i(\omega)$ s [19]:

$$\Pr[\bar{A}_1(\omega) \cap \bar{A}_2(\omega) \cap \dots \cap \bar{A}_k(\omega)] \geq \Pr[(\bar{A}_1(\omega))^k]. \quad (15)$$

From (14) and (15), we obtain:

$$\Pr[\text{good}|k, \omega] \leq 1 - \Pr[(\bar{A}_1(\omega))^k] \quad (16)$$

$$\leq 1 - (1 - \sum_{j=1}^k f_j)^k. \quad (17)$$

To derive (17), we used the fact that the probability of event $A_1(\omega)$ is maximized if the k most popular files is in ω . The obtained upper bound in (17) does not depend on ω . Hence,

$$\Pr[\text{good}|k] \leq 1 - (1 - \sum_{j=1}^k f_j)^k. \quad (18)$$

In the following, we will consider four non overlapping regions for $r(n)$ and for each region, we will prove the first part of the theorem.

D. First region

We first consider the region $r(n) = O(\sqrt{\frac{1}{n}})$. From (13) and (18),

$$E[L] \leq \frac{2}{r(n)^2} \sum_{k=0}^n [1 - (1 - \sum_{j=1}^k f_j)^k] \Pr[K = k] \quad (19a)$$

$$\leq \frac{2}{r(n)^2} \sum_{k=1}^n k \Pr[K = k] \sum_{j=1}^k f_j. \quad (19b)$$

Using part (iii) of lemma 1, the second summation $\sum_{j=1}^k f_j \leq 2 \frac{k^{1-\gamma_r}}{m^{1-\gamma_r}}$. Thus,

$$\begin{aligned} E[L] &\leq \frac{4}{r(n)^2} \sum_{k=0}^n \frac{k^{2-\gamma_r}}{m^{1-\gamma_r}} \Pr[K = k] \\ &\leq \frac{4}{r(n)^2 m^{1-\gamma_r}} \sum_{k=0}^n k^2 \Pr[K = k] \\ &= \frac{4}{r(n)^2 m^{1-\gamma_r}} E[K^2]. \end{aligned}$$

For the Binomial random variable $K = B(n, \alpha r(n)^2)$,

$$E[K^2] = (\alpha n r(n)^2)^2 + \alpha n r(n)^2 (1 - \alpha r(n)^2)$$

Therefore,

$$\begin{aligned} E[L] &\leq \frac{4}{r(n)^2 m^{1-\gamma_r}} ((\alpha n r(n)^2)^2 + \alpha n r(n)^2 (1 - \alpha r(n)^2)) \\ &= \frac{4n}{m^{1-\gamma_r}} (\alpha^2 n r(n)^2 + \alpha(1 - \alpha r(n)^2)) \\ &= c \frac{n}{m^{1-\gamma_r}}. \end{aligned}$$

where c is some constant.

E. Second region

Then, we consider the region that $r(n) = \Omega(\sqrt{\frac{1}{n}})$, and $r(n) = O(\sqrt{\frac{\log(m)}{n}})$. Equation (19a) implies:

$$\begin{aligned} E[L] &\leq \frac{2}{r(n)^2} \sum_{0 \leq k < k_0} [1 - (1 - \sum_{j=1}^k f_j)^k] \Pr[K = k] \\ &\quad + \frac{2}{r(n)^2} \sum_{k \geq k_0} [1 - (1 - \sum_{j=1}^k f_j)^k] \Pr[K = k]. \end{aligned} \quad (21)$$

Assuming that $r(n) \leq \sqrt{\frac{c \log(m)}{n}}$, we choose $k_0 = 6\alpha c \log(m)$ where c is some constant. Note $[1 - (1 - \sum_{j=1}^k f_j)^k]$ is an increasing function of k and it is less and equal to 1. Therefore, (21) implies,

$$\begin{aligned} E[L] &\leq \frac{2}{r(n)^2} [1 - (1 - \sum_{j=1}^{k_0} f_j)^{k_0}] \Pr[K < k_0] \\ &\quad + \frac{2}{r(n)^2} \Pr[K \geq k_0], \end{aligned} \quad (22)$$

$$\begin{aligned} &\leq \frac{2}{r(n)^2} [k_0 \sum_{j=1}^{k_0} f_j] \Pr[K < k_0] + \frac{2}{r(n)^2} \Pr[K \geq k_0] \\ &\leq \frac{4}{r(n)^2} \left[\frac{k_0^{2-\gamma_r}}{m^{1-\gamma_r}} \right] \Pr[K < k_0] + \frac{2}{r(n)^2} \Pr[K \geq k_0] \end{aligned} \quad (23)$$

$$(24)$$

We use part (iii) of lemma 1 to derive the last equation. For the binomial random variable K and for any $R \geq 6E[K]$, the Chernoff bound holds [18]:

$$\Pr[K \geq R] \leq 2^{-R}. \quad (25)$$

Applying the Chernoff bound and substituting k_0 in (24), we acquire:

$$\begin{aligned} E[L] &\leq \frac{4}{r(n)^2} \frac{(6\alpha c \log(m))^{2-\gamma_r}}{m^{1-\gamma_r}} \\ &\quad + \frac{2}{r(n)^2} 2^{-6\alpha c \log(m)} \\ &= 4(6\alpha c)^{2-\gamma_r} \frac{1}{r(n)^2} \frac{(\log(m))^{2-\gamma_r}}{m^{1-\gamma_r}} + \frac{1}{r(n)^2} \frac{2}{m^{6\alpha c \log 2}}. \end{aligned} \quad (26)$$

The function $f(x) = \frac{\log(x)}{x^\beta}$ is always less than $\frac{1}{\beta}$ where $\beta > 0$. Thus, $\log(m) \leq \frac{m^{\frac{\gamma_r}{2}}}{\eta^2}$.

$$\begin{aligned} E[L] &\leq 4(6\alpha c)^{2-\gamma_r} \frac{1}{r(n)^2} \left(\frac{m^{\frac{\gamma_r}{2}}}{\eta^2} \right)^{2-\gamma_r} \frac{1}{m^{1-\gamma_r}} + \frac{2n}{m^\eta} \\ &= \frac{4(6\alpha c)^{2-\gamma_r}}{\eta^{4-2\gamma_r}} \frac{1}{r(n)^2 m^\eta} + \frac{2n}{m^\eta} \\ &= \Theta\left(\frac{n}{m^\eta}\right) \end{aligned} \quad (27)$$

F. Third and fourth regions

For the third region, $r(n) = \Omega(\sqrt{\frac{\log(m)}{n}})$ and $r(n) = O(\sqrt{\frac{1}{n}})$. To show the upper bound for $E[L]$ in this region, we follow similar procedure in the second region by setting $k_0 = 6\alpha n r(n)^2$. For the last region $r(n) = \Omega(\sqrt{\frac{m^\eta}{n}})$, the total number of all virtual clusters $\frac{2}{r(n)^2} = O(\frac{n}{m^\eta})$. Thus, for this range of $r(n)$, $E[L] = O(\frac{n}{m^\eta})$.

In the following, we will show the *second part* of the theorem. Similar to proof of the theorem 1, we relate the number of good clusters and active links. We restrict users to communicate with their neighbors in their clusters. We further limit users not to get certain files from their neighbors although some neighbors might store these files. In this case the value of a cluster is the sum of probability of stored files that users can get via their neighbors. By the restriction on files the value of cluster becomes concentrated around its mean. We also consider self requests in finding the lower bound. Applying Chernoff bound and Azuma inequality we show that the probability of goodness is not vanishing when a user is surrounded in average by $\pi n r_{opt}(n)^2$ neighbors from which the result follows.

Define $\eta_1 \triangleq \eta + \epsilon = \frac{(1-\gamma_r)\gamma_c}{1-\gamma_r+\gamma_c}$. We should show that if we choose $r(n) = \Theta(\sqrt{\frac{m^{\eta_1}}{n}})$, the probability that a virtual cluster is good does not vanish as n grows.

When $r(n) = \Theta(\sqrt{\frac{m^{\eta_1}}{n}})$, there are $\Theta(\frac{n}{m^{\eta_1}})$ virtual clusters. The number of active D2D links is upper bounded by the number of virtual clusters. Thus, $E[L] = O(\frac{n}{m^{\eta_1}})$. Then, we show that for $c_3 \sqrt{\frac{m^{\eta_1}}{n}} \leq r(n) \leq c_4 \sqrt{\frac{m^{\eta_1}}{n}}$, $E[L] = \Omega(\frac{n}{m^{\eta_1}})$. To do this, we follow similar procedure in theorem 1. We divide the cell into virtual clusters and we allow each user to look for its desired file just within its cluster. As mentioned before, each cluster can block at most 16 other clusters (Figure 4(b)).

G. Limiting users and excluding self request

To find the lower bound, we even more restrict users. We assume that users can not get files $\{1, 2, \dots, q-1\}$ locally even if there are users in the cluster that cache these files where $q = m^{\frac{\eta_1}{\gamma_c}}$. So, caching files $\{1, 2, \dots, q-1\}$ doesn't have any value for any user in the cluster.

$E[L]$ is lower bounded by expression in (2) where the lower bound for $E[G]$ is given in (4). Similar to (6), we exclude

the self requests. Thus, the probability that a cluster is good conditioned on k and ω is

$$\Pr[\text{good}|k, \omega] \geq 1 - \left(1 - \left(v(\omega) - \max_{i \in \{q, \dots, m\}} f_{\omega_i}\right)\right)^k \quad (28)$$

where $v(\omega) = \sum_{j=q}^m f_j \mathbf{1}_j$ and $\mathbf{1}_j$ is an indicator function. $\mathbf{1}_j$ is one if at least one user in the virtual cluster stores file j . We limit ourselves to all cases in which at least one user caches file q . Hence,

$$\Pr[\text{good}|k, \omega] \geq 1 - (1 - (v(\omega) - f_q))^k \quad (29)$$

H. Chernoff bound

As in proof of theorem 1, we first limit the interval of k and then we use the Chernoff bound.

By restricting k to an interval around its average, i.e., $I = [nr(n)^2(1 - \delta)/2, nr(n)^2(1 + \delta)/2]$ where $0 < \delta < 1$ and applying (29) in (4), the following lower-bound is obtained:

$$\begin{aligned} E[G] &\geq \frac{2}{r(n)^2} \sum_{k \in I} \Pr[K = k] \\ &\quad \times \sum_{\omega \in \mathbf{x}} [1 - (1 - (v(\omega) - f_q))^k] \Pr[\omega], \end{aligned} \quad (30)$$

where $\mathbf{x} = \{\omega \mid |\omega| = k \text{ and } q \in \omega\}$. Let k^* be

$$k^* \triangleq \arg \min_{k \in I} \sum_{\omega \in \mathbf{x}} [1 - (1 - (v(\omega) - f_q))^k] \Pr[\omega]$$

Notice that k^* and also all $k \in I$ are $\Theta(m^{\eta_1})$. Then,

$$\begin{aligned} E[G] &\geq \frac{2}{r(n)^2} \Pr[K \in I] \\ &\quad \times \sum_{\omega \in \mathbf{x}} 1 - (1 - (v(\omega) - f_q))^{k^*} \Pr[\omega] \\ &\geq \frac{2}{r(n)^2} (1 - 2 \exp(-nr(n)^2 \delta^2 / 6)) \\ &\quad \times \sum_{\omega \in \mathbf{x}} [1 - (1 - (v(\omega) - f_q))^{k^*}] \Pr[\omega]. \end{aligned} \quad (31)$$

We use the Chernoff bound in (31). Let $A_{q,h}^{k^*}$ denote the event that $1 \leq h \leq k^*$ users cache file q . Then, we can rewrite the above lower-bound as follows:

$$\begin{aligned} E[G] &\geq \frac{2}{r(n)^2} (1 - 2 \exp(-nr(n)^2 \delta^2 / 6)) \\ &\quad \times \sum_{h=1}^{k^*} E_v[1 - (1 - (v - f_q))^{k^*} | A_{q,h}^{k^*}] \Pr[A_{q,h}^{k^*}] \\ &\geq \frac{2}{r(n)^2} (1 - 2 \exp(-nr(n)^2 \delta^2 / 6)) \\ &\quad \times \sum_{h=k^* p_q(1-\delta_1)}^{k^* p_q(1+\delta_1)} E_v[1 - (1 - (v - f_q))^{k^*} | A_{q,h}^{k^*}] \Pr[A_{q,h}^{k^*}] \end{aligned} \quad (32)$$

where $\Pr[A_{q,h}^{k^*}] = \binom{k^*}{h} (p_q)^h (1 - p_q)^{k^* - h}$, $k^* p_q$ is the average of binomial random variable h and $0 < \delta_1 < 1$.

Define h^* as

$$\begin{aligned} h^* &\triangleq \arg \min_{k^* p_q(1-\delta_1) \leq h \leq k^* p_q(1+\delta_1)} \\ E_v[1 - (1 - (v - f_q))^{k^*} | A_{q,h}^{k^*}] \end{aligned} \quad (33)$$

Using Chernoff bound for binomial random variable h , we obtain:

$$\begin{aligned} E[G] &\geq \frac{2}{r(n)^2} (1 - 2 \exp(-nr(n)^2 \delta^2 / 6)) \\ &\quad \times (1 - 2 \exp(-k^* p_q \delta_1^2 / 3)) E_v[1 - (1 - (v - f_q))^{k^*} | A_{q,h^*}^{k^*}] \end{aligned} \quad (34)$$

The probability that a user caches file q is:

$$\begin{aligned} p_q &= \frac{\frac{1}{q^{\gamma_c}}}{\sum_{j=1}^m \frac{1}{j^{\gamma_c}}} \\ &= \frac{\Theta(\frac{1}{m^{\eta_1}})}{H(\gamma_c, 1, m)} \end{aligned} \quad (35)$$

where function H is defined in lemma 1. We show in lemma 1 that $H(\gamma_c, 1, m) = \Theta(1)$ given that $\gamma_c > 1$. Thus, all $h \in [k^* p_q(1 - \delta_1), k^* p_q(1 + \delta_1)]$ are $\Theta(1)$. By selecting the constant c_3 large enough, the second exponential term $(1 - 2 \exp(-k^* p_q \delta_1^2 / 3))$ will be greater than zero.

I. Probability of goodness is not vanishing

To complete the proof, it is enough to show that the probability that cluster is good, i.e., $E_v[1 - (1 - (v - f_q))^{k^*} | A_{q,h^*}^{k^*}]$ given in (34) does not vanish.

$$\begin{aligned} E_v[1 - (1 - (v - f_q))^{k^*} | A_{q,h^*}^{k^*}] &\geq \\ \int_{|v - E_v[v | A_{q,h^*}^{k^*}]| < t} (1 - (1 - (v - f_q))^{k^*}) f_{v | A_{q,h^*}^{k^*}}(v) dv \end{aligned} \quad (36)$$

where $f_{v | A_{q,h^*}^{k^*}}(v)$ is a probability distribution function of value v conditioned on $A_{q,h^*}^{k^*}$ and $0 < t < E_v[v | A_{q,h^*}^{k^*}]$. The average of v conditioned on $A_{q,h^*}^{k^*}$ is given by:

$$E_v[v | A_{q,h^*}^{k^*}] = f_q + \sum_{j=q+1}^m f_j (1 - (1 - p_j)^{k^* - h^*}) \quad (37)$$

From equation (36) and since $(1 - (1 - (v - f_q))^{k^*})$ is an increasing function of v ,

$$\begin{aligned} E_v[1 - (1 - (v - f_q))^{k^*} | A_{q,h^*}^{k^*}] &\geq \\ &\geq (1 - (1 - ((E_v[v | A_{q,h^*}^{k^*}] - t) - f_q))^{k^*}) \\ &\quad \times \Pr[|v - E_v[v | A_{q,h^*}^{k^*}]| < t] \end{aligned} \quad (38a)$$

$$\begin{aligned} &\geq 1 - \exp(-k^* (E_v[v | A_{q,h^*}^{k^*}] - t - f_q)) \\ &\quad \times \Pr[|v - E_v[v | A_{q,h^*}^{k^*}]| < t] \end{aligned} \quad (38b)$$

We show in lemma 2, $E_v[v | A_{q,h^*}^{k^*}] = \Theta(\frac{1}{m^{\eta_1}})$. Thus, $k^* E_v[v | A_{q,h^*}^{k^*}] = \Theta(1)$. Furthermore, (36) implies

$$t = O(E_v[v | A_{q,h^*}^{k^*}]) = O(\frac{1}{m^{\eta_1}}).$$

Similar to (35), we can show that $f_q = \Theta(\frac{1}{m^{\eta_2}}) = O(E_v[v | A_{q,h^*}^{k^*}])$ where $\eta_2 = \frac{(1-\gamma_r)(1+\gamma_e)}{1-\gamma_r+\gamma_e}$. Thus, the exponent

in the first term of (38b) is $\Theta(1)$. To prove the result, it is enough to show the second term in (38b) does not approach zero as n grows. By applying the Azuma Hoeffding inequality in lemma 3,

$$\Pr[|v - E_v[v|A_{q,h^*}^{k^*}]| \leq t] \geq 1 - 2 \exp\left(-\frac{2t^2}{(k^* - h^*)(f_q)^2}\right) \quad (39)$$

Due to the fact that $k^* = \Theta(m^{\eta_1})$ and $h^* = \Theta(1)$, the term $k^* - h^* = \Theta(m^{\eta_1})$. If we select $t = \Theta(\frac{1}{m^{\eta_1}})$, we can observe that the exponent $\frac{2t^2}{(k^* - h^*)(f_q)^2}$ scales with $m^{\eta_2(2-\gamma_c)}$. Hence, if $\gamma_c < 2$, the exponent goes to infinity as n grows. $\gamma_c < 2$ implies that $\epsilon < \frac{1}{6}$. This means that v is concentrated around its average with high probability if $\epsilon \leq \frac{1}{6}$ and as a result, the second term in (38b) is positive constant when n goes to infinity.

APPENDIX C PROOF OF THEOREM 3

The proof of the first part of the theorem is similar to the proof of the theorem 2. $E[L]$ is upper bounded by the expression in (19a). Next, we consider three non-overlapping regions for $r(n)$ and we show the upper bound is valid for every $r(n)$.

A. First region

First, we assume $r(n) = O(\sqrt{\frac{\log \log(m)}{n}})$. Equation (19b) and part (v) of lemma 1 imply,

$$E[L] \leq \frac{2}{r(n)^2} \sum_{k=0}^n \frac{\log(k) + 1}{\log(m)} \Pr[K = k] \quad (40a)$$

$$\leq \frac{2}{r(n)^2 \log(m)} \sum_{k=0}^n k^2 \Pr[K = k] \quad (40b)$$

$$= \frac{2}{r(n)^2 \log(m)} E[K^2] \quad (40c)$$

$$\leq \frac{2}{r(n)^2 \log(m)} [(\alpha n r(n)^2)^2 + \alpha n r(n)^2] \quad (40d)$$

$$= \frac{2n}{\log(m)} [\alpha^2 n r(n)^2 + \alpha] \quad (40e)$$

$$\leq \frac{2cn}{\log(m)} (\alpha^2 \log \log(m) + \alpha) \quad (40f)$$

$$= \Theta\left(\frac{n \log \log(m)}{\log(m)}\right) \quad (40g)$$

To derive (40f), we use the range of $r(n)$.

B. Second and third regions

Let's consider the second region for $r(n)$. In this region $r(n) = \Omega(\sqrt{\frac{\log \log(m)}{n}})$ and $r(n) = O(\sqrt{\frac{\log \log(m)}{n}})$. From (19a),

$$\begin{aligned} E[L] &\leq \frac{2}{r(n)^2} \sum_{k=0}^{6\alpha n r(n)^2} \left[1 - \left(1 - \sum_{j=1}^k f_j\right)^k\right] \Pr[K = k] \\ &\quad + \frac{2}{r(n)^2} \sum_{k=6\alpha n r(n)^2}^n \left[1 - \left(1 - \sum_{j=1}^k f_j\right)^k\right] \Pr[K = k] \end{aligned} \quad (41)$$

where α is defined in theorem 2. The term $[1 - (1 - \sum_{j=1}^k f_j)^k]$ is an increasing function of k , thus,

$$\begin{aligned} E[L] &\leq \frac{2}{r(n)^2} \left[1 - \left(1 - \sum_{j=1}^{6\alpha n r(n)^2} f_j\right)^{6\alpha n r(n)^2}\right] \\ &\quad + \frac{2}{r(n)^2} \Pr[K > 6\alpha n r(n)^2] \end{aligned} \quad (42a)$$

$$\leq \frac{2}{r(n)^2} 6\alpha n r(n)^2 \sum_{j=1}^{6\alpha n r(n)^2} f_j + \frac{2}{r(n)^2} 2^{-6\alpha n r(n)^2} \quad (42b)$$

In (42b), we applied the Chernoff bound [18]. From lemma 1 and the range of $r(n)$, we obtain

$$\begin{aligned} E[L] &\leq 12\alpha n \frac{\log(6\alpha n r(n)^2) + 1}{\log(m)} + \frac{2}{r(n)^2} 2^{-6\alpha n r(n)^2} \\ &\leq 12\alpha n \frac{\log(6\alpha c_7 \log(m)) + 1}{\log(m)} \\ &\quad + \frac{2n}{c_8 \log \log(m)} 2^{-6\alpha c_8 \log \log(m)} \\ &= \Theta\left(\frac{n \log \log(m)}{\log(m)}\right) + \frac{2n}{c_8 \log \log(m)} \times \frac{1}{\log(m)^{6\alpha c_8 \log(2)}} \\ &= \Theta\left(\frac{n \log \log(m)}{\log(m)}\right). \end{aligned}$$

For the last region, i.e., $r(n) = \Omega(\sqrt{\frac{\log \log(m)}{n}})$, the total number of virtual clusters is $O(\frac{n}{\log(m)})$ and as a result, $E[L] = O(\frac{n}{\log(m)}) = O(\frac{n \log \log(m)}{\log(m)})$.

In the following, we will show the *second part* of the theorem. We propose a centralized algorithm that can match the upper bound. The BS divides the cell into virtual cluster of size $r(n) = \Theta(\sqrt{\frac{\log \log(m)}{n \log \log(m)}})$. Given that there are k users in a cluster, each of them should cache one of the k most popular files. We show that under this caching policy, we can match the upper bound. To find the lower bound, we assume that users can just find their desired files just within clusters they belong to. The lower bound for $E[L]$ and $E[G]$ are respectively given in (2) and (3). Limiting the range of k results in

$$E[G] \geq \frac{2}{r(n)^2} \sum_{k \in I} \Pr[\text{good}|k] \Pr[K = k], \quad (43)$$

where $I = [nr(n)^2(1-\delta)/2, nr(n)^2(1-\delta)/2]$. Under this centralized caching policy the value of stored files within a cluster with k users is $v(k) = \sum_{j=1}^k f_j$. The cluster is good if at least one user within a cluster requests one of the k most popular files not stored in its own cache

$$\begin{aligned}
\Pr[\text{good}|k] &\geq 1 - (1 - (v(\omega) - f_1))^k \\
&= 1 - \left(1 - \sum_{j=2}^k f_j\right)^k \\
&\geq 1 - \exp\left(-k \sum_{j=2}^k f_j\right) \\
&\geq 1 - \exp\left(-\frac{k(\log(k) - 1)}{\log(m) + 1}\right) \quad (44)
\end{aligned}$$

In the last equation we used lemma 1. The expression in (44) is an increasing function of k . Thus, (44) and (43) imply

$$\begin{aligned}
E[G] &\geq \left(1 - \exp\left(-\frac{k_{\min}(\log(k_{\min}) - 1)}{\log(m) + 1}\right)\right) \Pr[K \in I] \\
&\geq \left(1 - \exp\left(-c \frac{k_{\min} \log(k_{\min})}{\log(m)}\right)\right) \\
&\times \left(1 - 2 \exp\left(-nr(n)^2 \frac{\delta^2}{6}\right)\right) \quad (46)
\end{aligned}$$

where $k_{\min} = nr(n)^2(1 - \delta)/2 = \Theta(\frac{\log(m)}{\log \log(m)})$. We use the Chernoff bound to derive (46). As n grows, the second term in (46) goes to 1. It can be seen that the first term in (46) is also $\Theta(1)$. Thus, $E[G]$ and consequently $E[L]$ are $\Theta(\frac{n \log \log(m)}{\log m})$.

APPENDIX D

SOME PRELIMINARY LEMMAS

Lemma 1: i) If $\gamma > 1$ and $a = o(b)$, $H(\gamma, a, b) = \Theta(\frac{1}{a^{\gamma-1}})$.
ii) If $\gamma < 1$, $a = o(b)$, and $a = \Theta(1)$, $H(\gamma, a, b) = \Theta(b^{1-\gamma})$.
iii) if $\gamma_r < 1$, $\sum_{j=1}^k f_j \leq 2 \frac{k^{1-\gamma_r}}{m^{1-\gamma_r}}$.
iv) If $\gamma_c, \gamma_r > 1$, $\sum_{i=2}^m f_i p_i = \Theta(1)$.
v) if $\gamma = 1$, $\sum_{j=l}^k f_j \leq \frac{\log(k)+1}{\log(m)}$ and $\sum_{j=2}^k f_j \geq \frac{\log(k)-1}{\log(m)+1}$.
where $H(\gamma, a, b) = \sum_{j=a}^b \frac{1}{j^\gamma}$,

$$p_i = \frac{\frac{1}{i^{\gamma_c}}}{\sum_{j=1}^m \frac{1}{j^{\gamma_c}}}, \quad 1 \leq i \leq m. \quad (47)$$

and f_i is defined in (1).

Proof: We first prove the parts (i) and (ii) of the lemma. $\frac{1}{x^\gamma}$ is monotonically decreasing. Thus,

$$H(\gamma, a, b) \geq \int_{x=a}^b \frac{1}{x^\gamma} = \frac{b^{(-\gamma+1)} - a^{(-\gamma+1)}}{-\gamma + 1} \quad (48)$$

We also have the following inequality:

$$\begin{aligned}
H(\gamma, a, b) - \frac{1}{a^\gamma} &= \sum_{j=a+1}^b \frac{1}{j^\gamma} \\
&\leq \int_{x=a}^b \frac{1}{x^\gamma} = \frac{b^{(-\gamma+1)} - a^{(-\gamma+1)}}{-\gamma + 1} \quad (49)
\end{aligned}$$

Thus, $H(\gamma, a, b)$ satisfies:

$$\frac{b^{(-\gamma+1)} - a^{(-\gamma+1)}}{-\gamma + 1} \leq H(\gamma, a, b) \leq \frac{b^{(-\gamma+1)} - a^{(-\gamma+1)}}{-\gamma + 1} + \frac{1}{a^\gamma} \quad (50)$$

Therefore, if $\gamma > 1$, $H(\gamma, a, b) = \Theta(\frac{1}{a^{\gamma-1}})$. Besides, if $\gamma < 1$ and $a = \Theta(1)$, then $H(\gamma, a, b) = \Theta(b^{1-\gamma})$.

For part (iii), using (48) and (49), we have

$$\begin{aligned}
\sum_{j=1}^k f_j &= \frac{H(\gamma_r, 1, k)}{H(\gamma_r, 1, m)} \\
&\leq \frac{k^{(1-\gamma_r)} - \gamma_r}{m^{(1-\gamma_r)} - 1} \\
&\leq 2 \frac{k^{(1-\gamma_r)}}{m^{(1-\gamma_r)}}.
\end{aligned}$$

Next we show part (iv). From (1), we have:

$$\begin{aligned}
\sum_{j=2}^m f_j p_j &= \frac{\sum_{j=2}^m \frac{1}{j^{\gamma_r + \gamma_c}}}{\sum_{j=1}^m \frac{1}{j^{\gamma_r}} \sum_{j=1}^m \frac{1}{j^{\gamma_c}}} \\
&= \frac{H(\gamma_c + \gamma_r, 2, m)}{H(\gamma_c, 1, m) H(\gamma_r, 1, m)} \quad (51)
\end{aligned}$$

When $\gamma_c, \gamma_r > 1$, both the nominator and the dominator of $\sum_{j=2}^m f_j p_j$ are $\Theta(1)$, from which (iv) follows.

Since the proof of the part (v) is similar to parts (i) and (ii), we omit it. ■

Lemma 2: If $\gamma_c > 1$, $\gamma_r < 1$, $k = \Theta(m^{\eta_1})$, and $h = \Theta(1)$

$$E_v[v|A_{q,h}^k] = \Theta(\frac{1}{m^{\eta_1}})$$

where $\eta_1 = \frac{\gamma_c(1-\gamma_r)}{1-\gamma_r+\gamma_c}$, $q = m^{\frac{\eta_1}{\gamma_c}}$, and $E_v[v|A_{q,h}^k]$ is defined in (37).

Proof: For the lower-bound, we have:

$$\begin{aligned}
E_v[v|A_{q,h}^k] &= f_q + \sum_{j=q}^m f_j (1 - (1 - p_j)^{k-h}) \\
&\geq \sum_{j=q}^m f_j (1 - e^{-k' p_j}) \quad (52)
\end{aligned}$$

where $k' = k - h = \Theta(m^{\eta_1})$. Using the taylor series, we obtain:

$$\begin{aligned}
E_v[v|A_{q,h}^k] &\geq \sum_{j=q}^m f_j k' p_j + f_j \frac{1}{2!} (k' p_j)^2 + f_j \frac{1}{3!} (k' p_j)^3 + \dots \\
&= k' \frac{H(\gamma_c + \gamma_r, q, m)}{H(\gamma_r, 1, m) H(\gamma_c, 1, m)} + \frac{1}{2!} k'^2 \frac{H(2\gamma_c + \gamma_r, q, m)}{H(\gamma_r, 1, m) H(\gamma_c, 1, m)^2} \\
&+ \frac{1}{3!} k'^3 \frac{H(3\gamma_c + \gamma_r, q, m)}{H(\gamma_r, 1, m) H(\gamma_c, 1, m)^3} + \dots \quad (53)
\end{aligned}$$

Parts (i) and (ii) of lemma 1 imply that all terms in the above equation are $\Theta(\frac{1}{m^{\eta_1}})$.

For showing the upper bound,

$$\begin{aligned}
E_v[v|A_{q,h}^k] &= f_q + \sum_{j=q+1}^m f_j(1 - (1 - p_j)^k) \\
&\leq f_q + k \sum_{j=q}^m f_j p_j \quad (54) \\
&\leq \frac{1}{q^{\gamma_r} H(\gamma_r, 1, m)} + k \frac{H(\gamma_c + \gamma_r, q, m)}{H(\gamma_r, 1, m) H(\gamma_c, 1, m)} \quad (55)
\end{aligned}$$

If we apply the results of lemma 1, we can show that $E_v[v|A_{q,h}^k]$ is $O(\frac{1}{m^{\eta_1}})$. ■

Lemma 3: For $t < E_v[v|A_{q,h}^k]$,

$$\Pr[|v - E_v[v|A_{q,h}^k]| \leq t] \geq 1 - 2 \exp\left(-\frac{2t^2}{(k-h)(f_q)^2}\right) \quad (56)$$

Proof: Function $v : \{1, 2, \dots, m\}^k \rightarrow R$ is equal to

$$v(\omega_1, \omega_2, \dots, \omega_k) = \sum_{i \in \tilde{\omega} \cap Q} f_i$$

where ω_j is the file that user j stores, $\tilde{\omega} = \bigcup_{j=1}^k \omega_j$ and $Q = \{q, q+1, \dots, m\}$. v is the sum of popularity of union of files stored by users when only files in set Q are considered to be valuable. By replacing the i th coordinate ω_i by some other value the value of v can change at most by f_q , i.e.,

$$\sup_{\omega_1, \dots, \omega_k, \hat{\omega}_i} |v(\omega_1, \dots, \omega_k) - v(\omega_1, \dots, \hat{\omega}_i, \omega_{i+1}, \dots, \omega_k)| \leq f_q$$

Using Azuma-Hoeffding inequality [20],

$$\Pr[|v - E_v[v|A_{q,h}^k]| \geq t] \leq 2 \exp\left(-\frac{2t^2}{(k-h)(f_q)^2}\right) \quad (57)$$

the result follows. ■

REFERENCES

- [1] "http://www.cisco.com/en/us/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html."
- [2] V. Chandrasekhar, J. Andrews, and A. Gatherer, "Femtocell networks: a survey," *Communications Magazine, IEEE*, vol. 46, no. 9, pp. 59–67, 2008.
- [3] N. Golrezaei, K. Shanmugam, A. Dimakis, A. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *INFOCOM*. IEEE, 2012.
- [4] —, "Wireless video content delivery through coded distributed caching," in *ICC*. IEEE, 2012.
- [5] N. Golrezaei, A. Molisch, and A. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *Accepted in ICC'12 WS - ViOpt*.
- [6] N. Golrezaei, A. Molisch, A. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *Accepted in IEEE Communications Magazine*, 2012.
- [7] D. A. Golrezaei, N. and A. Molisch, "Device to device transmission for increasing video throughput in wireless networks," *To be submitted for publication*.
- [8] P. Gupta and P. Kumar, "The capacity of wireless networks," *Information Theory, IEEE Transactions on*, vol. 46, no. 2, pp. 388–404, 2000.
- [9] A. Ozgur, O. Lévêque, and D. Tse, "Hierarchical cooperation achieves linear capacity scaling in ad hoc networks," in *INFOCOM 2007. 26th IEEE International Conference on Computer Communications*. IEEE, 2007, pp. 382–390.

- [10] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad-hoc wireless networks," in *INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 3. IEEE, 2001, pp. 1360–1369.
- [11] M. Franceschetti, M. Migliore, and P. Minero, "The capacity of wireless networks: information-theoretic and physical limits," *Information Theory, IEEE Transactions on*, vol. 55, no. 8, pp. 3413–3424, 2009.
- [12] "http://traces.cs.umass.edu/index.php/network/network."
- [13] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 1–14.
- [14] M. Penrose and O. U. Press, *Random geometric graphs*. Oxford University Press Oxford, 2003, vol. 5.
- [15] A. Molisch, *Wireless communications*. Wiley, 2011.
- [16] E. Lawler, J. Lenstra, A. Kan, and E. U. E. Institute, "Generating all maximal independent sets: Np-hardness and polynomial-time algorithms," *SIAM J. Comput.*, vol. 9, no. 3, pp. 558–565, 1980.
- [17] Y. Chen, C. Caramanis, and S. Shakkottai, "On file sharing over a wireless social network," in *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 249–253.
- [18] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *The Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493–507, 1952.
- [19] R. Holley, "Remarks on the fkg inequalities," *Communications in Mathematical Physics*, vol. 36, no. 3, pp. 227–231, 1974.
- [20] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge Univ Pr, 2005.